
Journal of Informatics and Web Engineering

Vol. 4 No. 2 (June 2025)

eISSN: 2821-370X

Diabetes Risk Prediction using Shapley Additive Explanations for Feature Engineering

Chinwe Miracle Chituru¹, Sin-Ban Ho^{2*}, Ian Chai³

^{1,2,3} Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia

*corresponding author: (sbho@mmu.edu.my, ORCID: 0000-0003-2995-2120)

Abstract - Diabetes is prevalent globally, expected to increase in the next few years. This includes people with different types of diabetes including type 1 diabetes and type 2 diabetes. There are several causes for the increase: dietary decisions and lack of exercise as the main ones. This global health challenge calls for effective prediction and early management of the disease. This research focuses on the decision tree algorithm utilization to predict the risk of diabetes and model interpretability with the integration of SHapley Additive exPlanations (SHAP) for feature engineering. Random forest and gradient boosting models were developed to identify the risk factors and compare the prediction with the decision tree model. The performance of these classifiers was evaluated using the metrics for accuracy, f1-score, precision, and recall. Understanding the features that drive predictions can enhance clinical decision-making as much as predictive accuracy. With the use of a comprehensive dataset having 520 instances with 17 features including the target output, the proposed decision tree model had an accuracy of 97%. The decision tree model's categorical variables enable straightforward data visualization. The SHAP tool was applied to interpret the model's prediction after developing the model. This is crucial for healthcare practitioners as it provides specific health metrics to identify high-risk diabetic patients. Preliminary results indicate that a combination of polyuria, polydipsia, and age are predictors of diabetes risk. This study highlights the benefits that the integration of SHAP and decision trees algorithm provides predictive capability and transparent model interpretability. It also contributes to the growing body of literature on machine learning in the healthcare industry. The results advocate for the application of this methodology in clinical settings for prediction fostering trust between the approach and practitioners and patients alike.

Keywords— Diabetes Risk Prediction, Decision Tree Algorithm, Additive Explanations, Feature Engineering, Data Visualization

Received: 22 November 2024; Accepted: 14 January 2025; Published: 16 June 2025

This is an open access article under the [CC BY-NC-ND 4.0](#) license.



1. INTRODUCTION

Diabetes is among the pressing health challenges faced in recent times and it affects millions of individuals around the world. The disease causes underlying issues in humans and makes people susceptible to other diseases or health challenges. Diabetes is rapidly spreading in adults, and children are also affected. When carbohydrate-rich foods are consumed, our body converts the nutrients to glucose or blood sugar which is sent into the blood all through the body

[1]. Glucose is an important metabolic fuel for the brain. Since glucose circulates through the blood, blood cells use glucose as their energy source. The hormone needed by the body to keep the range of the blood glucose level normal, *insulin* moves glucose from our bloodstream into cells in the body cells to make energy. Once insulin attaches itself

to the cell surface, glucose can enter the cell from the blood. It can be stored as energy for later use or immediately used for crucial body sustenance. Without insulin, glucose does not reach the cells, so glucose levels become higher than normal; this is called hyperglycemia. Diabetes is a result of the body failing to secrete adequate insulin or effectively use the insulin produced, thus excessively raising blood sugar levels. Early detection and management are essential.

Globally, a WHO-driven initiative unites stakeholders in combating diabetes to reduce the risk and ensure access to complete, affordable care and prevention [2]. The goal of predicting early diagnosis is important in the healthcare sector. With the development of machine learning techniques over the years, the predictive analytics landscape in health has completely changed. Among them, decision trees and random forests classifiers have gained significance due to their intuitiveness and interpretability [3], their value in clinical settings is great because transparency in this field plays an important role. Models have been developed for prediction, classification, regression, and hand gesture recognition. Tan et al. [4] proposed an SDViT which refers to the Stacking of Distilled Vision Transformers model, and this model achieved a notable accuracy of 100.00% for hand gesture recognition. Gradient boosting has been seen to fix errors from previous trees and provide better performance. To analyse the features in models SHAP values have been used to measure the contribution of the features to the model.

Given the worldwide burden of diabetes [2] and the need for accurate yet interpretable predictive tools [1], [3], this study seeks to enhance the usability of tree-based algorithms in healthcare. Additionally, SHAP can be used for interpretability [3], [5] by explaining feature contributions to the model's predictions. With the influx of machine learning models, these models have now become much more complex [4] and better trained to give accurate predictions, this makes it even harder to comprehend the explanation of the process behind the models' prediction. The increasing trend has resulted in the emergence of a growing interest in machine learning model explainability [5]. The widely used tool for this process is SHAPley values [5] which measure the impact of each feature and its contribution towards the prediction [6]. SHAP values enable one to see which features are most important to the model and how they impact the outcome. This study explores the integration of the decision tree algorithm and its efficiency in predicting diabetes with feature engineering, interpretability [6], and explainability by the SHAP-based method. In this connection, we hope to reveal some significant predictors of diabetes, to help enhance early diagnostic capability, and to contribute to the elaboration of improved preventive measures. In this regard, we would like to be hopeful that the present study fills an important gap between complex data engineering and practical clinical applications to fill a niche in data-driven diabetes management.

This study focuses on predicting and analysing diabetes risk using machine learning algorithms, with an emphasis on model interpretability and explainability using SHAPley values. Most feature variables in diabetes risk prediction models - such as body mass index, obesity, and age - are derived from patient data that require medical testing in clinical settings by health practitioners. Ensuring the accuracy of input data during patient surveys is also essential for reliable results. The dataset used in this study, the Sylhet Diabetes Hospital dataset (SDHD), includes 520 patient records, representing a local sample rather than a global one, which may limit the model's generalizability. SHAP values were applied to evaluate feature importance and assess the impact of each feature on model predictions, enhancing our understanding of which factors most significantly influence diabetes risk.

2. LITERATURE REVIEW

Prediction of diabetes risk has gained much importance in recent decades, and the invention of machine learning techniques recently has accelerated it further. Among these techniques, decision trees have gained much popularity due to their interpretability and ability to handle complicated datasets effectively. Nipa et al. [6] applied thirty-five classifiers on a merged dataset derived from the SDHD and pre-diagnosis diabetes dataset (PDD) and found that the Extra tree (ET) outperformed other classifiers with a 97.11% accuracy, while Multi-layer Perceptron, had an accuracy of 96.42%. In the same study, Hist Gradient Boosting Classifier (HGBC) and Light Gradient Boosting Machine (LGBM) had the maximum accuracy of 94.90% for the combined datasets. Saboor et al. [7] applied an optimised selector on the SDHD, followed by data normalization, preprocessing, and model building, and after evaluation, the decision tree model had the highest accuracy of 90.10%, while deep learning had an accuracy of 89.06%, the random forest had an accuracy of 88.02%, and gradient boosting had the accuracy of 89.58%. Jiang et al. [8] used machine learning techniques to assess the degree of connection between diabetes risk factors using specified feature variables. They found that optimising questionnaire variables and question count can greatly increase efficiency for accurate diabetes prevention while maintaining accuracy. The study found that obesity was the variable that most strongly

influences the risk of diabetes [8]. Other predictor attributes such as psychological problems, high cholesterol, alcohol abuse, coronary heart disease, and low family income have less of an influence in predicting diabetes than obesity. Faniqul et al. [9] utilised different data classification algorithms to find the one with the best accuracy in predicting early-stage diabetes and concluded that random forest has the highest accuracy of 97.4%.

2.1. Decision Trees in Diabetes Risk Prediction

Decision trees have been applied for diabetes risk prediction in several studies that have shown the potential of this approach in determining important risk factors. Mahmoud et al. [10] implemented a task scheduling algorithm; the decision tree technique is among the popular machine learning algorithms for building and visualizing predictive models. Saboor et al. [7] applied a decision tree model to analyse patient data provided by the Sylhet Diabetes Hospital in Sylhet, Bangladesh dataset and found that age, obesity, and gender, among other features, are major predictors of the onset of diabetes. Another useful study by Azad et al. [11] developed combined prediction models using the Synthetic Minority Over-sampling TEchnique (SMOTE), decision tree algorithm, and a genetic algorithm for diabetes classification that achieved a high accuracy of 82.1256%, further depicting the enriched predictive power among machine learning methods. Research from Dritsas and Trigka [12] on diabetes risk prediction using the SDHD dataset found that random forest and KNN were the most successful models achieving 98.59% accuracy, with the SMOTE with 10-fold cross-validation. Additionally, these models reached 99.22% accuracy when SMOTE was applied with an 80:20 train-test split.

The decision tree is a simple model that for prediction, splits data into branches based on features. Decision trees clearly outline the decision criteria for regression and classification trees [10]. A decision tree also has leaf nodes and decision nodes. Decision nodes split the data into branches based on the value of the feature, and each node represents a condition or question that splits data based on an attribute to help further classify the data. Leaf nodes are terminal nodes that provide the final prediction based on the data that reaches them. Random forests combine many decision trees to make more accurate predictions. In gradient boosting, the models build trees sequentially to improve performance by correcting mistakes of the previously built trees – these trees are a structure that makes predictions based on input data. While decision trees come with many advantages such as the ability to model nonlinear relationships, interpretability, and ease of understanding, it has several shortfalls. The main problem concerning decision trees is overfitting, significantly when using small data sets for model training. Methods such as pruning and ensemble techniques have addressed the issue of overfitting while maintaining predictive accuracy. Random forest, for example, is an ensemble method that builds multiple decision trees using bootstrapped samples [6] and random feature selection, which reduces overfitting and improves model accuracy. Bootstrapped samples involve randomly selecting data points, allowing for repetition, which helps create diverse trees. Gradient boosting, another ensemble method, constructs decision trees in sequence, with each new tree correcting the errors of the previous one. This process enhances model robustness and improves overall predictive performance.

Although decision tree models are simpler and more interpretable [6], [8]. They remain a valuable tool for understanding decision-making processes, as they allow users to intuitively visualise how decisions are made. Compared to ensemble methods like random forests and gradient boosting, decision trees are less computationally expensive and faster to train, making them ideal for initial analyses and simple applications.

2.2. SHAPley Explanations in Feature Interpretability

SHAPley values is a theoretical game method to explain the features of a machine learning model's output. SHAP values can be used to analyse samples and measure important attributes for interpreting model results globally [3]. SHAPley values are model agnostic and thus the explanations can apply to any model, as opposed to model specific that only work for interpretation. The SHAP tool is applied after the model has been trained.

The dataset was split into 2 sets; to test and train the model. 80% of the data was used to train the model and 20% to test the model. Ge et al. [13] cross-evaluated deep learning models for volatility using 70% for training, 15% for testing, and 15% for evaluation, 70-15-15 train-validation-test split. This approach was utilised because it did not violate the temporal aspect of the data. With more reviews on similar research for binary classification problems and classifiers on segmenting the training and dataset, this study measures the performance using the 80-20 train-test split

[14], [15], with the training data having a larger portion. Empirical studies show that 20% to 30% of the data used for testing to gauge the accuracy of the model produces the best results.

SHAP has probably become the most important method for feature interpretability in machine learning models, drawing its root from cooperative game theory. Using Shapley values, SHAP provides a systematic approach to attribute the contributions of individual features to model predictions, gaining better insight and trust in ML systems. In theoretical foundations, SHAP embodies a game theoretical approach, whereby the Shapley value, named after Lloyd Shapley, who introduced it in 1953, provides a fair method of payment to the players (features) regarding their marginal contribution to the total outcome. The reason this approach is unique lies in the fact that it treats problems of fairness and consistency for feature importance attribution problems [3], particularly when complex models are involved, and other methods of interpretability fail. Machine Learning Applications contains recent works that have demonstrated that SHAP can be applied to a wide array of domains, including healthcare and finance. These methods have been used to explain deep learning model predictions of disease diagnosis in patients, providing insight into which features are most informative to clinicians. An example is symptoms versus laboratory test results. In finance, SHAP has also been applied to credit scoring to help explain how an individual's attributes, such as income or credit history, are used in determining whether to approve a loan.

SHAP offers several advantages over other interpretability techniques, such as LIME. While Local Interpretable Model Agnostic Explanations (LIME), provide local approximations, SHAP guarantees consistency and is globally interpretable [3], offering a more comprehensive understanding of features. Both LIME and SHAP techniques allow for personalised risk factor explanations and capture non-linear relationships between input features [16]. However, SHAP stands out as a technique applicable to any model, from simple linear regressions to complex ensemble methods, which broadens its utility across various applications. One of SHAP's key features is its powerful visualization tools such as force plots and summary plots which help clarify feature contributions. These visualizations not only enhance interpretability but also facilitate communication with stakeholders, promoting transparency in Artificial Intelligence (AI) decision-making.

Despite its strengths, SHAP is not without challenges and limitations. For example, explaining interactions between features can be difficult, which complicates the interpretation of complex relationships. Additionally, as observed in this study, SHAP values can exhibit slight fluctuations in feature importance when the model is run multiple times. The computational cost of calculating Shapley values can also be high, making it challenging to apply to large datasets or complex models.

Despite the progress of decision trees in diabetes prediction, there is a need for further detailed investigations involving broadened populations and global datasets. It is also necessary to track the trends over time using longitudinal data. Current research often focuses on demographics, and gathering responses across different populations across the globe is limited due to these specifics. Additionally, there is no consensus as to which feature selection methods will serve best, which may be critical for the performance of machine learning models like decision trees, random forests, and gradient boosting.

3. RESEARCH METHODOLOGY

Analysis was carried out using the dataset from the Hospital in Sylhet, Bangladesh, and collected by Kaggle's collaborator [17] in 2020 with reference to the research of Faniqul et al. [9] on the prediction of diabetes at an early stage using mining techniques. For this paper, a clean dataset of 520 patient records out of which 320 records have positive (Class 1) results, and 200 instances have negative (Class 0) results. The dataset was preprocessed to separate numerical and categorical features and then encoded categorical variables using the label encoder class imported from Python Scikit-learn library. The decision tree machine learning algorithm, a widely used supervised learning method for binary classification tasks was utilised for prediction. Using the Seaborn python library, the correlation matrix tool was implemented to visualise how the variables related to each other. To ensure improved performance, hyperparameter tuning involving grid search cross-validation having wider range values for each hyperparameter helped find the best configuration of the model. The GridsearchCV class was imported from the model selection module in Scikit-learn. Ensemble methods: gradient boosting and random forest were applied for boosting and bagging to allow the combination of weak learners to perform as strong learners and reduce variance.

Prediction and Explainability using Decision Trees and SHAP proceeds with creating a tuned decision tree model on the dataset first for accuracy. Then, after the training, and testing, SHAP: SHapley Additive exPlanations, a technique

for explaining the model's predictions, is utilised. SHAP values provide insights into how each feature contributes to the outcome, thus providing ultimate clarity on the decision-making processes. Shap Tree Explainer is initialised with the trained model to calculate the SHAP values. The tree explainer is an exact method that is optimised to explain SHAP values for tree-based models. The input features include age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity.

The dataset has 16 feature variables, and the target variable Diabetes has 2 classes: 0 for Negative and 1 for Positive. They were classified under three categories, namely, demographics, symptoms, and clinical signs. After careful consideration of diabetes with the contributing and exacerbating factors, no feature selection by dropping columns after retrieving the dataset was done, and the description of the feature variables used to train and test the model is highlighted in Table 1, Table 2, and Table 3. The feature variables listed in Table 1 describes the demographic of the respondents. The table presents demographic data relating to diabetic and non-diabetic occurrences. It is divided into three segments: age, gender, and obesity. Each row specifies the dataset detail.

Table 1. Key Input Features for Demographics

Demographics		
Feature	Dataset Detail	Description
Age	Numerical value	Minimum age = 16 Maximum Age = 90
Gender	Male / Female	Male = 63% Female = 37%
Obesity	Yes/No	Excess body fat

The clinical symptoms features are indicated in Table 2 with key insights on polyuria, polydipsia, polyphagia, sudden weight loss, weakness, genital thrush, visual blurring, and delayed healing. All are nominal types as they have distinct categories and no inherent order. The features selected from this table help to analyse the patterns and correlation between the health indicators listed and the prevalence of diabetes.

Table 2. Key Input Features Selection for Clinical Symptoms

Clinical Symptoms			
Feature	Type	Binary	Description
Polyuria	Nominal	Yes/No	Excessive urination
Polydipsia	Nominal	Yes/No	Increased thirst
Polyphagia	Nominal	Yes/No	Extreme hunger
Sudden weight loss	Nominal	Yes/No	Unexplained rapid loss of body weight
Weakness	Nominal	Yes/No	Decreased physical strength
Genital thrush	Nominal	Yes/No	Fungal yeast infection
Visual blurring	Nominal	Yes/No	Distorted vision
Delayed healing	Nominal	Yes/No	Troubled recovery from wounds

The key input features of the lifestyle factors are explained in Table 3. The following table summarizes the lifestyle factors related to diabetes frequency. Physical and psychological factors' input features are itching, irritability, partial

paresis, muscle stiffness, and alopecia. It also gives a brief description and all the features in Table 3 are selected to train the model.

Table 3. Key Input Features for Clinical Signs

Physical and Psychological signs			
Feature	Type	Binary	Description
Itching	Nominal	Yes/No	Unpleasant sensation on the skin; desire to scratch
Irritability	Nominal	Yes/No	Increased annoyance or sensitivity
Partial paresis	Nominal	Yes/No	Muscle weakness in part of the body
Muscle stiffness	Nominal	Yes/No	Reduced flexibility, tightness in muscle
Alopecia	Nominal	Yes/No	Hair loss

3.1. Approach for Visualization Distributions

Data visualization can be done to represent the information given graphically in a visual context. Visual representation and design are important, especially in statistical quantities [18]. Various elements like charts maps, graphs, plots, and charts can be used to identify the patterns and insights from the data. To prepare the data and generate the heatmap, the data was organised in a comma-separated values (CSV) file. Each cell in the file corresponds to a relative value. The data was processed for normalization. The categorical and numerical values were separated, and then a copy of the data was created to avoid DataFrame modification. Using the sklearn python library, preprocessing module, Onehotencoder, label encoder, and Min-max scaler were imported. The Min-max scaler was to normalize the numerical values and the encoders to encode the categorical features to the format in machine-readable format. However, for the data visualization, the original pre-normalised dataset was used for intuitive visualization.

3.2. Decision Tree Prediction Model

The decision tree starts as a single node and branches into possible outcomes of the variable node [19], which makes it an important classification algorithm [3], [19]. The model was built using the decision tree algorithm. The dataset was split into sets to test and train the model. 80% of the data was used to train the model and 20% to test the model. Figure 1 shows the architecture with patient data preprocessing, feature engineering in scaling and normalisation.

The DataFrame (*df*) shape has 520 records and 17 columns. After dropping the target variable column, the feature variables have 16 columns left to train and test the model. The dataset has no missing values. Feature variables are marked as *X* in the study; *X* has 16 columns, and the target is marked as *y*, and *y* has 1 column. The summation of the shapes of the train and test sets for *X* and *y* was 416 and 104 to tally with the percentage value for each set.

The classification report of the decision tree model had a precision accuracy of 97.00 % (0.97). This study focuses on solving a binary classification, so there are 2 outcome classes - Class 0 and Class 1. The *Negative class 0* has a precision of 0.97 and an f1-score of 0.97 while the *Positive Diabetes class 1* showed a precision of 0.99 and an f1-score of 0.99. Hyperparameters improve ML models' accuracy immensely, the configuration was set before training the model. The *k*-fold cross-validation via the Grid search technique was used to find the best hyperparameter values tuning. With *cv* = 5, the training data was split into 5 parts to ensure a reliable model evaluation and robust tuning.

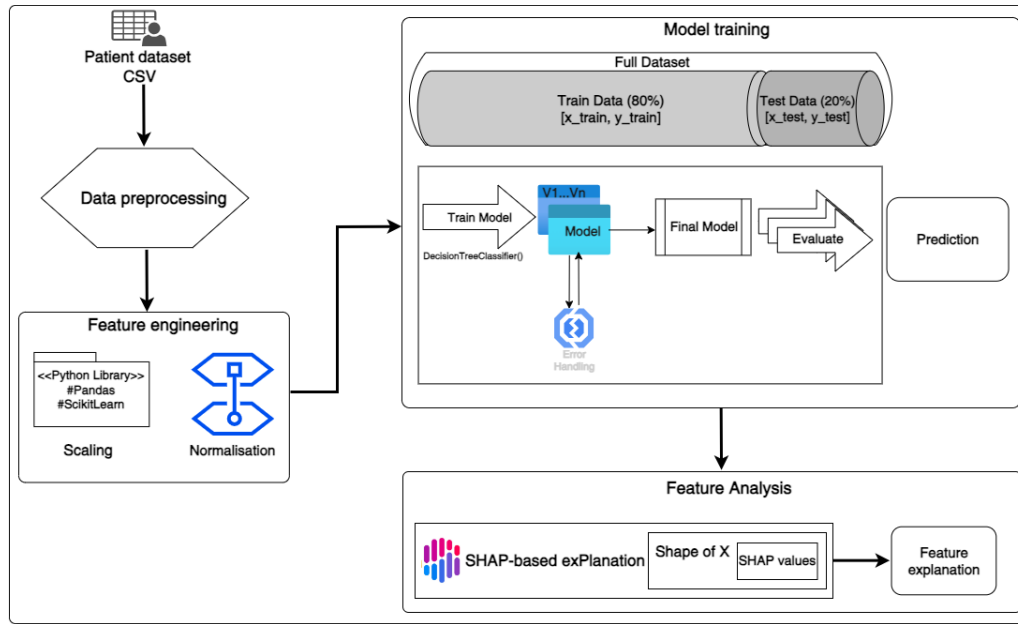


Figure 1. Illustration of the 80-20 Train-test Split Evaluation Model

3.3. SHAPley Feature Engineering

SHAP values should have a shape that matches the data making it identical to the X shape. Utilizing the SHAP Python library and the SHAP Tree explainer with the Decision Tree Classifier tree-based model, the X values for this model are Shape of X (520, 16) indicating 520 samples and 16 features in the dataset. The shape of the X test is (104, 16). The Shape of SHAP values is (104, 16, 2), indicating 104 samples in the subset data to be explained using SHAPley values, 16 as the number of features, and 2 as the SHAP values for both classes 0 and 1. SHAP can be installed from the SHAP library. To apply SHAP, the values are obtained from the trained decision tree model and test data that was passed to the SHAP tree explainer model. SHAP visualisation tools such as `shap.summary_plot()`, `shap.force_plot()` and `shap.dependence_plot()` to interpret and visualize which features most influence predictions.

Data engineering techniques help users gain a better understanding of more meaningful insights and visualizations regarding the trends in the severity of the disease symptoms [20]. Data handling techniques guarantee the accuracy of data, where decision depends upon the integrity of every bit of information, accurate data engineering [21], and reliable decision-making processes can be efficient [22], [23]. Shapley values provide a framework that correctly attributes the contribution of every feature in a dataset to the overall prediction for decision reliability. They are helpful in the derivation of meaningful insights and drive data-driven strategy formulation [24] by quantifying variable importance. In addition, Shapley value visualization can present complex relationships in a more transparent and far-reaching manner, enabling stakeholders to visualize insights into more intuitive and actionable formats.

4. RESULTS AND DISCUSSIONS

4.1. Prediction Models for Diabetes and Pre-diabetes

Using classification models namely decision tree, random forest, and gradient boosting, all three machine learning (ML) classifier models had a high-test accuracy (97%–99%) and high Area Under the Curve (AUC) values from 0.97 to 1.00. The model with the highest accuracy was the random forest model (99%), with a sensitivity of 97%, and AUC of 1.00; its specificity (97%) was neither the highest nor lowest among the three models as shown in Table 4. In contrast, although the decision tree model had maximum accuracy (97%), sensitivity (97%), and AUC (1.00), its specificity (88%) was the lowest. The gradient boosting classifier gave a good accuracy (97%), sensitivity (95%),

specificity (100%), and AUC value (1.00). The key evaluation metrics for the classifiers are shown in Table 4. Overall, the predictive models for diabetes risk had similar and good prediction performance with only slight differences.

Table 4. Key Evaluation Metrics

Feature	Decision Tree	Random Forest	Gradient Boosting
Accuracy	97%	99%	97%
Precision (Class 0)	0.97	1.00	0.92
Precision (Class 1)	0.99	0.99	1.00
Recall (Class 0)	0.97	0.97	1.00
Recall (Class 1)	0.99	1.00	0.96
f1-score (Class 0)	0.97	0.99	0.96
f1-score (Class 1)	0.99	0.99	0.98
AUC value	0.97	1.00	1.00
Sensitivity (Recall)	0.97	1.00	0.95
Specificity	0.88	0.97	1.00

4.2. Visualization of Patterns and Distributions Post-Normalization

Using the Python Seaborn and Matplotlib libraries to analyse and create the plot, Figure 2 illustrates a correlation heatmap with each input variable shown by a row and column and the colour strength highlights the correlation strength, that is darker colours indicate a stronger correlation. The colour ranges from yellow (negative correlation) to dark blue (extremely strong positive correlation). The values in each cell show the degree of linear association between the variables in the dataset and are correlation coefficients (r) that range from -0.4 to 1.0. The correlation heatmap acts as a data engineering tool that shows the relationships among variables and is important for machine learning systems.

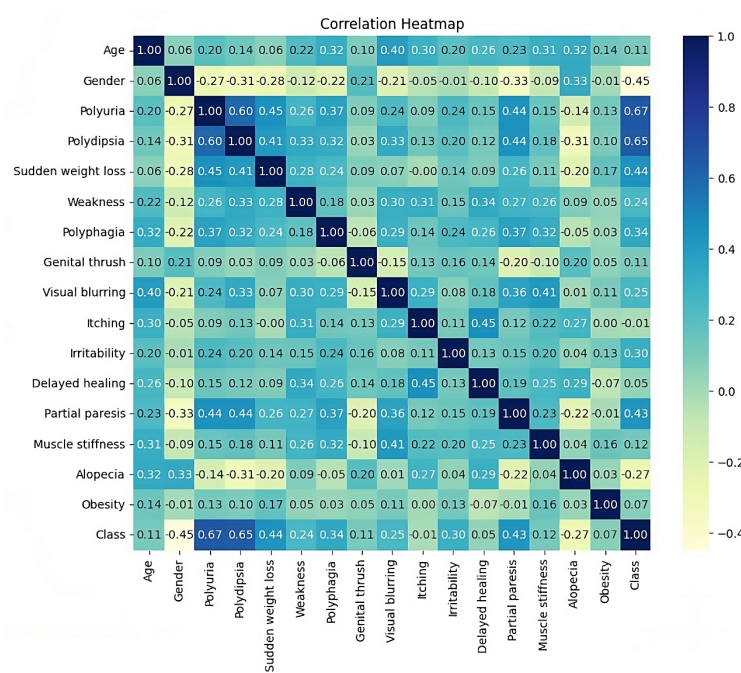


Figure 2. Correlation Heatmap between the Health-related Variables

The correlation interpretation involves the use of both the numerical values and the colours. From the above heatmap, class as the output variable shows significant correlations with the predictors. With a focus on class as the output variable, strong positive correlations can be seen between class and polyuria at 0.67, and with polydipsia at 0.65. Moderate positive correlations between class and sudden weight loss at 0.44, partial paresis at 0.43, polyphagia at 0.34, and irritability at 0.30. Weak positive correlations can be seen between visual blurring at 0.25, muscle stiffness at 0.12, genital thrush, and age having the same value of 0.11. Although these two variables have a weak correlation, they are significant and cannot be overlooked. Class has very weak or negligible correlations with obesity (0.07) and delayed healing has almost no correlation (0.05). Also, the heatmap shows a weak negative correlation with alopecia (-0.27) and gender (-0.45).

Regarding the relationship between predictors, polyuria and polydipsia ($r = 0.60$), polydipsia contributes strongly to polydipsia, but the r value is not greater than 0.7, so both variables can be included in the prediction model. This also implies that a patient with high polydipsia tends to have high polyuria. Moderate correlations in delayed healing and itching (0.45), polyuria, and sudden weight loss (0.45) indicate a positive but not particularly strong relationship. Weak positive correlations displayed in genital thrush and gender (0.21) indicate that both variables do not have significant linear associations. Visual blurring and gender, on the contrary, show a weak negative correlation (-0.21).

Alopecia refers to the partial or complete loss of hair. As such, we could observe that there is a weak perceptible correlation response between alopecia and gender from Figure 2. In such a fraction of time when this dataset is collected, the correlation heatmap visualization can be overcome with compassion and amazement for the gentle relationships being, which dared to observe the very correlations to provide such insights. Through this good correlation state, one could have a mind of wisdom to see things as they really are in attaining serenity, bliss, and further directions towards the subsequent cultivation of experiences.

Distributions of continuous features were compared against categorical values. Using the Python library Matplotlib and the *pyplot* sub-module, Figure 3 box and whisker plot shows the distribution of obesity by age and gender. From the input, the continuous feature is age as it takes a wide range of numerical values, and the categorical values are obesity and gender as they indicate a binary condition of “Yes/No” and “Male/Female” respectively.

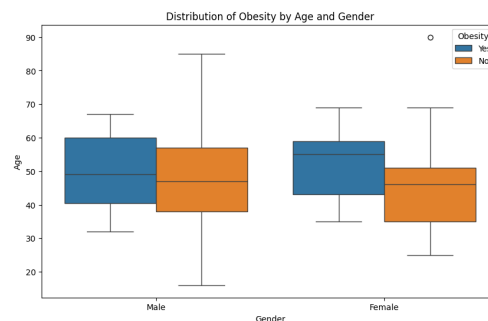


Figure 3. Obesity Prevalence by Age and Gender

A key observation from the box plot in Figure 3 is that the boxes with obesity which are blue in colour are higher on the scale, which infers that older people experience obesity. The median (black line in the box) is higher for females with obesity and mid-range for males. The median for the obesity group is 50 years for males and 55 years for females. The upper quartile of obesity indicates 60 years for males and has a slight contrast of 59 years for females. The no-obesity groups have longer whiskers demonstrating a broad age range. The whiskers are the vertical lines above and below each box excluding the outliers. Outliers are found at age 90 for females with no obesity. The dot above the whiskers represents the outliers in the no-obesity group. Overall, the box and whisker plot shows that obesity is more prevalent in older females than males.

4.3. Binary Decision Trees

Binary nodes in decision trees are split based on a condition and divided into two branches which can be called child nodes, and this is the tree model used to make predictions. Each node forms two branches for further splitting; a left

(true) branch and a right (false) branch. The supervised learning model used a graphical representation in a treelike flow structure. Figure 4 shows the decision tree visualization for this model with the gini value decreasing in the subsequent child nodes. The tree is structured hierarchically with each node having splits into branch nodes. Each node predicts the class based on the samples. For the split criterion, the Gini index is used to minimize impurity, maximize class accuracy, and decide on the best splits. Features in Figure 4 include variables such as polyuria, gender, age, polydipsia, alopecia, and weakness. The features from the dataset are not all illustrated here due to the large dimension of the full decision tree, so the internal nodes after the second level split and leaf nodes are hidden and shown as ellipses. The ellipses represent the continuation of the tree structure that is not shown for brevity.

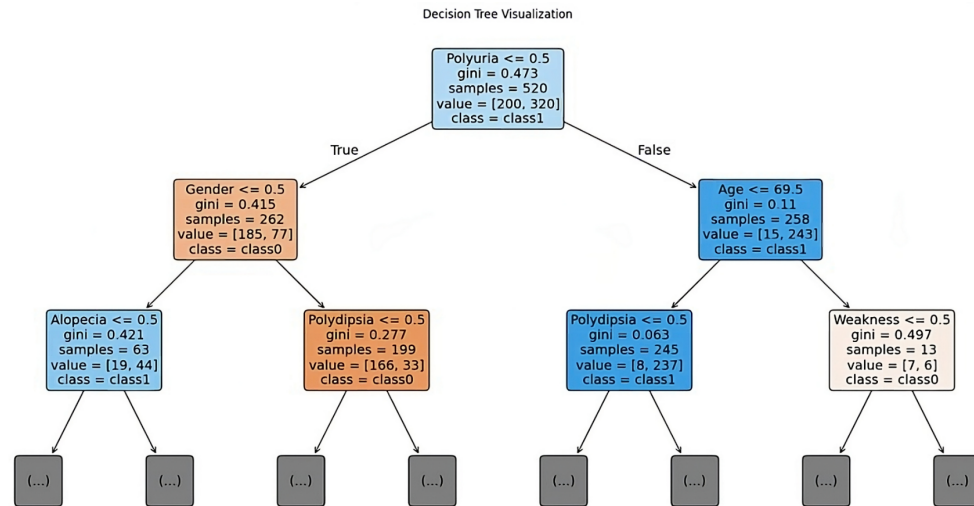


Figure 4. Decision Binary Tree Visualisation

The root node starts the entire dataset, having the polyuria feature (≤ 0.5). The 0.5 value is the threshold that the decision tree has used to split the data based on polyuria. All instances where polyuria is ≤ 0.5 go to the left and all instances where polyuria is > 0.5 , go to the right node. Gini (= 0.473) indicates the Gini measure before splitting. Samples = 520 refers to the total data points in the dataset and the value of [200, 320] is the distribution of class labels, where the numbers are the count of instances of each class in that node [class 0 and class 1] respectively. As highlighted in the research methodology section above, 320 records have positive (Class 1) results, and 200 instances have negative (Class 0) results. Since class 1 has many instances, it is the prediction for the root node. The root node further splits into the internal nodes which represent the decision points. Instances where polyuria ≤ 0.5 is shown at the left(true) branch and further splits on gender (≤ 0.5). The majority of the dataset falls to this branch (samples = 262) and the value = [185, 77], indicates that class 0 has higher instances in this node. The right (false) branch represents instances where polyuria (> 0.5) splits on age (≤ 69.5) and conversely, class 1 is the predicted class which is determined by the majority class from the value = [15, 243] of the 258 instances in this node.

The lower-level splits in Figure 4 show more decision points in the internal nodes. The split condition for the 63 samples from the first node on the left is alopecia (≤ 0.5). Gini impurity (0.421) indicates a moderate impurity and the values [19, 44] of the 63 instances show that 19 samples belong to the negative class while 44 belong to the positive class, giving the majority class to be class 1 since it has the majority of samples (44 out of 63). The second node from the left applies the split condition of (polydipsia ≤ 0.5) and contains 199 samples. A lower gini impurity (0.277) indicates a purer node compared the previous node. Of the 199 samples, 166 instances belong to class 0, and 33 belong to class 1, so this node predicts class 0 because it has the majority class.

Further split below the right (false) node in Figure 4, the third node from the left, polydipsia ≤ 0.5 is the split condition for 245 samples. This is another split on polydipsia highlighting its importance in the different classes. Gini impurity (0.063) implies that the node is almost pure, that is, most samples belong to a single class. The value [8, 237] confirms that class 1 is dominant (237 out of 245 samples) in this node, so this node predicts class 1. For the fourth node from the left, 13 instances are split based on weakness (≤ 0.5), and the gini impurity (0.497) is in proximity with the gini value (0.5) indicating that the classes are almost evenly mixed. 7 of the 13 samples belong to class 0 and 6 belong to

class 1. Although the split in this node is close, class 0 with 7 instances dominates and is the predicted class. In this lower-level split, alopecia and weakness are less influential compared to polydipsia.

4.4. SHAP Summary plot

The summary bar plot indicates the feature's importance. Figure 5 shows the mean SHAP values plot. As visualised, Polyuria has a large impact on the model's output, followed by polydipsia, gender, partial paresis, and age. Visual blurring does not affect the model. Partial paresis is in the fourth position in the bar plot but in the sixth position in the beeswarm and violin plot in Figure 6 indicating the difference when plotted separately for the individual classes.

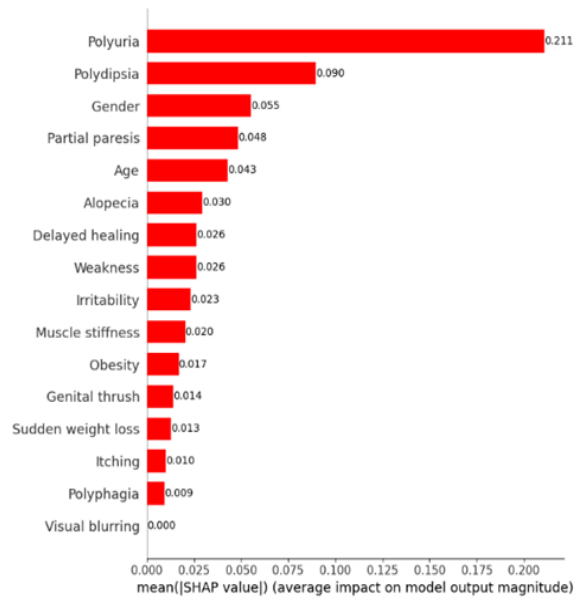


Figure 5. SHAP Summary Bar Plot

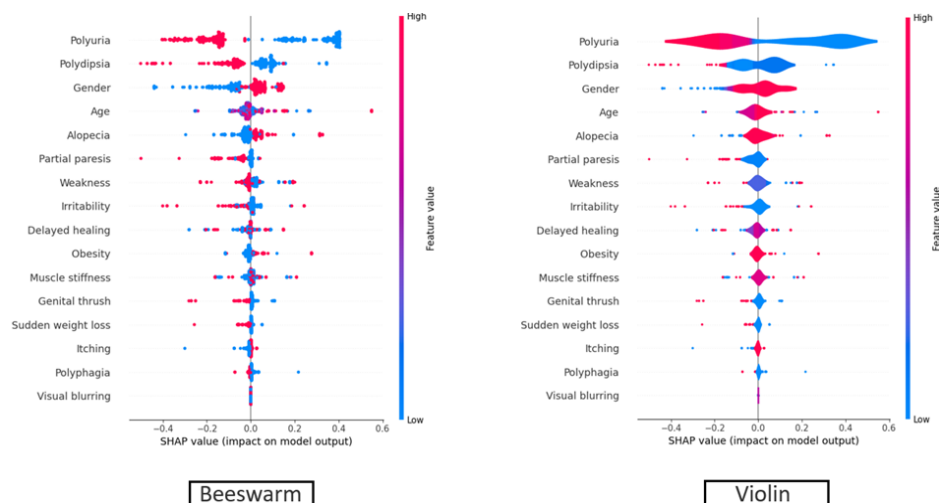


Figure 6. SHAP Summary Plot

The violin and beeswarm plots in Figure 6 were implemented to see the difference between the two classes, positive and negative, where the left side is the negative class and the right side is the positive class, plus the directionality impact. In these plots, the x-axis represents the SHAP value, and the y-axis shows all the features. The colour gradient

from blue to red represents the value of the features from low to high. Each point on the beeswarm plot is one SHAP value of an instance from the dataset. In Figure 6, polyuria is the value with the highest influence value.

Polyuria is inversely correlated, because the polyuria feature value increases, the SHAP impact value decreases, and as the polyuria feature value decreases the SHAP value increases. And feature values such as visual blurring, have negligible effects. The higher feature values of gender, age, and alopecia have a higher impact on the model. In contrast, the lower feature value of polydipsia has a higher impact on the model. This SHAP summary plot presents a detailed insight into feature contributions within the predictive model, for the health classification problem to predict diabetes. This plot depicts how each feature influences the model output, including but not limited to polyuria, polydipsia, gender, and age.

Key observations are that for polyuria and polydipsia, the extreme values of these features (those far to the left, represented by the red points) are strongly positive on the model output, indicating that when these symptoms are present, the likelihood of a positive diagnosis increases. The SHAP value distribution of gender depicts that gender differences have had a moderate influence on the model's predictions, though less pronounced than polyuria and polydipsia. Age and sudden weight loss SHAP values of the features are well distributed between positive and negative contributions, suggesting that the impact of this feature depends on other factors. Obesity and itching have a more mixed impact, demonstrating that they can influence the prediction in both directions, depending on the case. This analysis reflects that the model is sensitive to some clinical features. Symptoms such as polyuria and polydipsia were key predictors, while features like age and gender interacted in a complicated way, implying the need for careful consideration of feature impacts during interpretation.

4.5. SHAP Value Heatmap

The heatmap visualizes SHAP values for classes 0 and 1 across different features and samples. In Figure 7, the x-axis shows the model input features, while the y-axis represents instances, numbered from 0 to 100. The SHAP values are encoded on a colour scale, ranging from dark blue (lower values) to dark red (higher values), with white or light colours representing values close to zero. The samples are ordered based on a hierarchical clustering by their explanation similarity, resulting in samples with similar model output for the same reason being grouped together. For example, respondents with polyphagia, itching, and obesity are grouped in the plot of Figure 7.

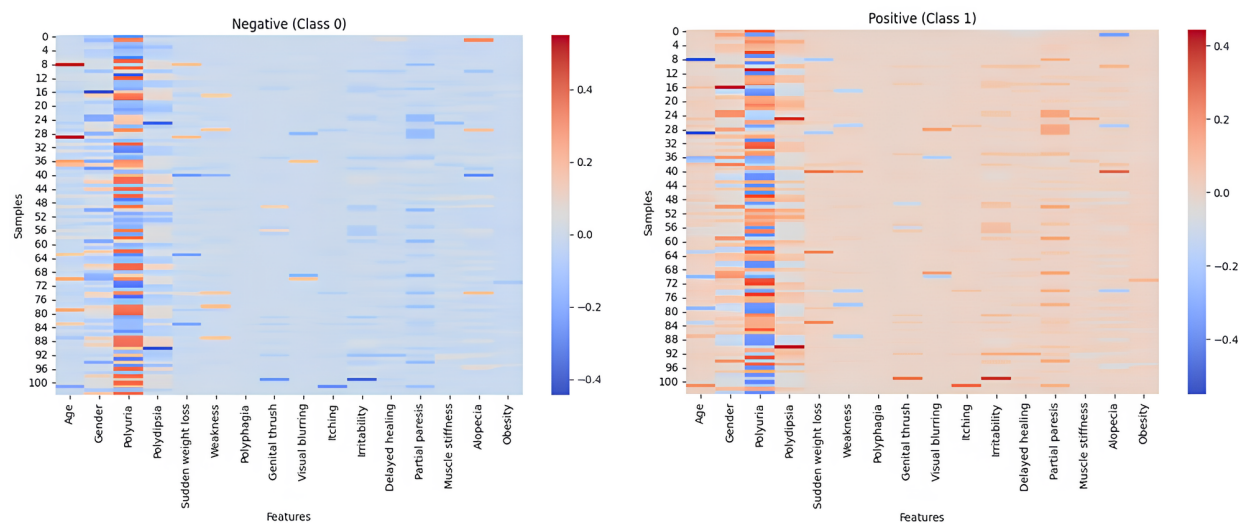


Figure 7. SHAP Values Heatmap for Classes 0 and 1

Red cells indicate features that positively contribute to the prediction of the classes, such as polyuria for both classes, age for class 0, and in contrast polydipsia and gender for class 1. Blue cells show features that negatively impact the prediction of classes, such as irritability for class 0 and age for class 1. Light or white cells suggest minimal impact on the prediction.

Notable patterns in classes 1 and 0 reveal that polyuria and polydipsia display strong red patterns for many samples, indicating that these features often strongly contribute to the prediction. In contrast, age, gender, and obesity exhibit more mixed patterns, suggesting variable impacts across different samples. The variability in feature impact is evident from the changing colours within each column. The heatmap is a valuable tool for understanding feature importance and model behaviour in machine learning, particularly for tasks such as diabetes risk assessment.

4.6. SHAP Force plot

The SHAP force plot illustrates the important features influencing the model prediction and is a valuable tool for error analysis. It visualizes the output for a single-instance prediction of the model, making it useful for explaining which features are most important for a specific prediction and how they contribute to the model's output. In this example, a test set sample was analysed to determine which input features contributed to the prediction of class 0 for the tenth instance in the dataset. This analysis, shown in the force plots of Figure 8, was rendered using the Matplotlib Python library.

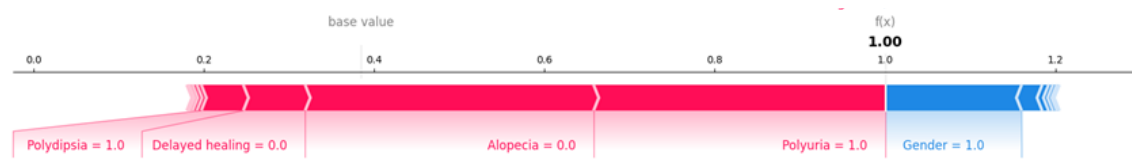


Figure 8. Force Plot Single Prediction in Class 0

The force plot provides details about the testing instance, including the expected value and SHAP values. The size of each feature in the plot represents its local magnitude, while the colour indicates the direction and strength of the impact: the red arrows pointing right represent features pushing the prediction higher than the base value, blue arrows pointing left represent features pushing it lower. For instance, alopecia and polyuria are shown to have a higher impact on the prediction. Gender, with a value of 1.0, contributes significantly to lowering the prediction.

In this case, the sum of these features' contributions results in a final prediction $f(x)$ of 1.00, where the positive and negative contributions balance each other. In Figure 8, polydipsia and polyuria, both with values of 1.0, slightly increase the prediction above the base value (0.38), while gender with a value of 1.0 reduces the prediction.

Figure 8 indicates that alopecia at 0.0, delayed healing at 0.0, polyuria at 1.0, and polydipsia at 1.0 have contributed to pushing the prediction higher. The base value of 0.38 represents the expected average model prediction over the training dataset before applying the feature contributions. The model output value in bold is the model's final prediction $f(x)$ at 1.00. Both these values are shown at the top of Figure 8.

To analyse a single prediction using the second instance of the dataset predicting class 1, Figures 8 and 9 highlight five features that contributed to the prediction for the tenth and second instances: polydipsia, delayed healing, alopecia, polyuria, and gender. Figure 9 illustrates the same features as Figure 8 but with opposite effects. Specifically, for the tenth instance in class 0, gender pushes the prediction lower toward 0, while for the second instance in class 1, gender pushes the prediction higher.

Regarding the magnitude of impact and the prediction, the $f(x)$ values of 1.00 and 0.00 at the top of the figures represent the final predictions for this instance. In Figure 9, polydipsia and polyuria, both with values of 1.0, are pushing the prediction lower, which contrasts with their behaviour in Figure 8, where they contribute to pushing the prediction higher.

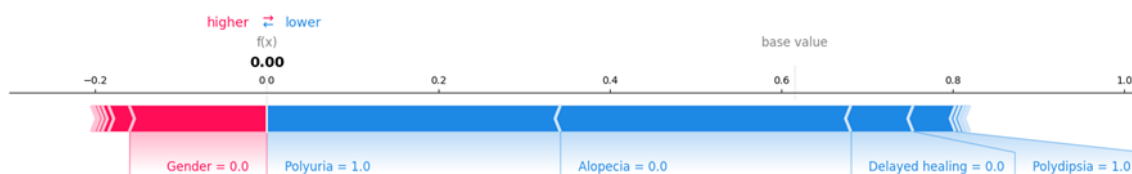


Figure 9. Force Plot Single Prediction in Class 1

4.7. Decision Plot

SHAP decision plots visualize how models make decisions, providing deep insight into model behaviour by connecting cumulative SHAP values for each prediction. Features are ordered by importance in descending order. The x-axis in Figure 10 represents the model output, ranging from -0.2 to 1.0, while the y-axis lists the model features. Each observation is represented by a coloured line corresponding to a specific prediction.

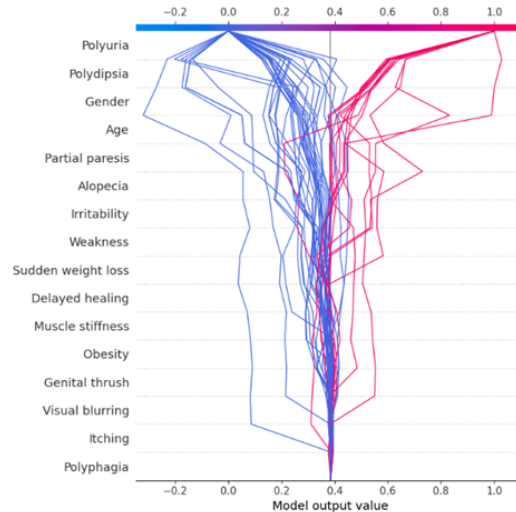


Figure 10. SHAP Decision Plot for Target Class 0

The plot structure is as follows: the y-axis lists the input features, the x-axis is the model output value (i.e. probability from -0.2 to 1.0), and each line represents a different instance or record, corresponding to a different respondent in the dataset. The gradient from blue to pink indicates the range of prediction values, where blue represents lower values and pink represents higher values. This gradient visually reflects the progression of predictions from the lowest to the highest.

In this study, 100 random instances from the x-test values were used for the decision plot. The decision plot in Figure 10 illustrates how each line contributes to the model's prediction for class 0 (no diabetes class) and highlights the features that influenced these predictions. Decision plots are capable of handling and visually displaying the impact of many features. In this case, the plot is tilted towards the values 0.0 and 1.0. It is read from the bottom to the top, where the final position of each line at the top represents the model's final prediction for that instance. Decision plots are intuitive because they provide a literal representation of SHAP values. Notably, the lines spread significantly at the features polyuria, polydipsia, and gender, which are identified as the most informative features, with polyuria being the most influential.

Figure 10 shows that polyuria and polydipsia have SHAP values that tilt towards both the negative and positive sides, indicating that their values can either decrease or increase the model output. Higher values for polyuria push the model output higher, while lower values push it lower. Features such as gender, age, partial paresis, and alopecia tend to have a relatively more positive influence on the output when their values are high. Conversely, features like obesity, genital thrush, and visual blurring exhibit more centralised effects, as indicated by the SHAP values for some instances.

Furthermore, the SHAP decision plot shows how different features contribute to model predictions across many instances for the decision plot with target class 1 (diabetes) in Figure 11. For this plot, the x-axis range is 0.0 to 1.4. In Figure 11, the plot is tilted towards 0.0 and 1.0, as observed in Figure 10. The features on the y-axis are typically ordered in their descending importance. This importance is calculated over the 100 observations that were plotted. This is different from the importance of the entire dataset. At the top, each line strikes at the corresponding x-axis. The features are ranked lower, such as polyphagia and visual blurring with less influence on the final prediction. For the class 1 feature importance in Figure 11, polyuria is the most informative feature, while gender and polydipsia are ranked as the next two most influential features. Alopecia, weakness, and sudden weight loss had an average impact.

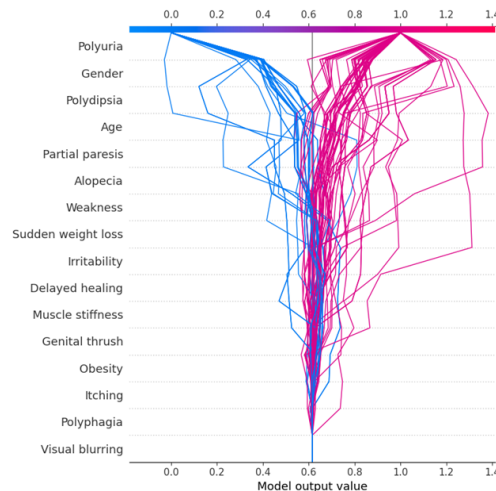


Figure 11. SHAP Decision Plot for Target Class 1 (Diabetes)

The pattern analysis reveals a cluster of cases with predictions jumping directly to high probabilities (> 0.8), while another cluster is observed in the middle ranges, between 0.4 to 0.6. The line spread demonstrates significant variation in predictions due to the combination of features. This plot is particularly useful for visualizing which features have the greatest impacts on the predictions and how their combinations lead to different outcomes. Additionally, it helps identify subgroups within the dataset that exhibit specific prediction trends, which may appear at the extremes – either with exceptionally good results or extremely poor results. This insight can be critical for understanding and improving model performance.

The ordering of features in the decision plot is designed to illustrate how the prediction flows from a base value to the final prediction. Additionally, the decision plot supports hierarchical cluster feature ordering and user-defined feature ordering, allowing for flexible analysis. In Figure 11, some features form distinct clusters on either side of the plot, indicating that polyuria, polydipsia, and gender are strongly associated with the model predictions. Notably, gender appears in the second position for class 1, whereas it ranks third for class 0. This highlights the differences in the importance of the gender feature explanation across the two classes. The relationships between features infer that, if more lines lean towards the higher values, for example, in pink, for polyuria and polydipsia, can suggest that these features are positively correlated with the model output. This indicates they are predictive of the diabetes condition that the model is trained to detect. Consequently, the decision plot provides an objective way to understand the relationships between different features and the model predictions.

4.8. Future Enhancements

While the findings of this study using the current model contribute significant details to predicting diabetes risk and feature interpretation, future enhancements could extend its impact. This model could be enhanced further to make more accurate and more interpretable predictions. The integration of real-time wearable data from continuous monitoring could be one such option [25], along with the expansion of the dataset to more diverse patient populations to increase model generalizability. For feature engineering, the SHAPley paradigm can be applied to a long short-term memory (LSTM) model trained on diabetes time series data as this data may provide insight into how risk factors change over time, which may lead to more personalised predictions [26]. Also, health systems with SHAP-based prediction models could enable healthcare providers to make informed, and data-driven decisions potentially to be integrated with knowledge and experience models such found in [27][28]. The way forward is that SHAPley can be used currently together with framework documentation [29][30] for the future enhancement.

5. CONCLUSION

Among the three classifier models assessed, the random forest demonstrated the best model performance. This model achieves high accuracy and an AUC value equal to that of the gradient boosting model. However, the random forest

was preferred for diabetes risk prediction due to its highest sensitivity in detection rate. For feature interpretability, the SHAP values were calculated using the decision tree model. SHAP engineering revealed that the clinical symptoms of polyuria and polydipsia are the most impactful features in the predictive model for making accurate predictions. These symptoms drive the model output in alignment with clinical knowledge about their relevance in diabetes prediction. In contrast, features like age and gender, while still useful in the model output, exhibit less variability across the population, indicating a more moderate influence on predictions.

In the context of SHAP, variability refers to the differences in the degree of influence that input features exert on the model predictions across different instances. In this paper, the authors highlight that the SHAP values effectively explain the model's decision-making process and reveal how each feature contributes to the predictions. This interpretability is crucial in an AI-driven health model, where building trust in patient care is concerned. To ensure reliability, additional model tuning and validation with diverse real-world data are necessary, especially for features with high impact variability. This step is vital for improving the robustness and generalizability of the model in real-world clinical settings.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their valuable comments.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Chinwe Miracle Chituru: Formal Analysis, Investigation, Conceptualization, Data Curation, Methodology, Validation, Visualization, Resources, Writing – Original Draft Preparation;

Sin-Ban Ho: Conceptualization, Validation, Visualization, Resources, Supervision, Writing – Original Draft Preparation;

Ian Chai: Project Administration, Supervision, Writing – Original Draft Preparation, Review & Editing

CONFLICT OF INTERESTS

No conflict of interests was disclosed.

ETHICS STATEMENTS

The dataset used for this study was obtained from Kaggle and it is a public domain.


REFERENCES

- [1] J. A. M. Rexie, P. Santhosh, P. N. Solomon, and P. A. Vishnu, "Early Prediction of Diabetes using Several Machine Learning Algorithms," in *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, May 17, 2023, pp. 449–453. doi: 10.1109/iciccs56967.2023.10142749.
- [2] E. W. Gregg et al., "Improving health outcomes of people with diabetes: target setting for the WHO Global Diabetes Compact," *The Lancet*, vol. 401, no. 10384, pp. 1302–1312, Mar. 2023, doi: 10.1016/s0140-6736(23)00001-6.
- [3] A. S. Antonini et al., "Machine Learning model interpretability using SHAP values: Application to Igneous Rock Classification task," *Applied Computing and Geosciences*, vol. 23, p. 100178, Sep. 2024, doi: 10.1016/j.acags.2024.100178.
- [4] C. K. Tan, K. M. Lim, C. P. Lee, R. K. Y. Chang, and A. Alqahtani, "SDVIT: Stacking of Distilled Vision Transformers for Hand Gesture Recognition," *Applied Sciences*, vol. 13, no. 22, p. 12204, Nov. 2023, doi: 10.3390/app132212204.

- [5] H. Ulutas, R. B. Günay, and M. E. Sahin, "Detecting diabetes in an ensemble model using a unique PSO-GWO hybrid approach to hyperparameter optimization," *Neural Computing and Applications*, Jul. 2024, doi: 10.1007/s00521-024-10160-y.
- [6] N. Nipa, M. H. Riyad, S. Satu, N. Waliullah, K. C. Howlader, and M. A. Moni, "Clinically adaptable machine learning model to identify early appreciable features of diabetes," *Intelligent Medicine*, vol. 4, no. 1, pp. 22–32, Feb. 2023, doi: 10.1016/j.imed.2023.01.003.
- [7] A. Saboor, A. U. Rehman, T. M. Ali, S. Javaid, and A. Nawaz, "An Applied Artificial Intelligence Technique For Early Prediction of Diabetes Disease," in *2022 Third International Conference on Latest Trends in Electrical Engineering and Computing Technologies (INTELLECT)*, Nov. 16, 2022, pp. 1–6. doi: 10.1109/intellect55495.2022.9969401.
- [8] L. Jiang et al., "A feature optimization study based on a diabetes risk questionnaire," *Frontiers in Public Health*, vol. 12, Feb. 2024, doi: 10.3389/fpubh.2024.1328353.
- [9] M. M. Faniqul, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in *Advances in Intelligent Systems and Computing*, 2020, vol. 992, pp. 113–125. doi: 10.1007/978-981-13-8798-2_12.
- [10] H. Mahmoud, M. Thabet, M. H. Khafagy, and F. A. Omara, "Multiobjective task scheduling in cloud environment using Decision tree algorithm," *IEEE Access*, vol. 10, pp. 36140–36151, Jan. 2022, doi: 10.1109/access.2022.3163273.
- [11] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," *Multimedia Systems*, vol. 28, no. 4, pp. 1289–1307, Jun. 2021, doi: 10.1007/s00530-021-00817-2.
- [12] E. Dritsas and M. Trigka, "Data-Driven Machine-Learning Methods for Diabetes Risk Prediction," *Sensors*, vol. 22, no. 14, p. 5304, Jul. 2022, doi: 10.3390/s22145304.
- [13] W. Ge, P. Lalbakhsh, L. Isai, A. Lensky, and H. Suominen, "Comparing deep learning models for the task of volatility prediction using multivariate data," *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2306.12446.
- [14] P. Ruediger-Flore, M. Klar, M. Hussong, A. Mukherjee, M. Glatt, and J. C. Aurich, "Comparing binary classification and autoencoders for Vision-Based anomaly detection in material flow," *Procedia CIRP*, vol. 121, pp. 138–143, Jan. 2024, doi: 10.1016/j.procir.2023.09.241.
- [15] N. Boyko, "Evaluating binary classification algorithms on data lakes using machine learning," *Revue D Intelligence Artificielle*, vol. 37, no. 6, pp. 1423–1434, Dec. 2023, doi: 10.18280/ria.370606.
- [16] R. O. Alabi, M. Elmusrati, I. Leivo, A. Almangush, and A. A. Makitie, "Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP," *Scientific Reports*, vol. 13, no. 1, Jun. 2023, doi: 10.1038/s41598-023-35795-0.
- [17] "Early-Stage Diabetes Risk Prediction Dataset," Kaggle, Sep. 21, 2020. <https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset/data>
- [18] J. Rogel-Salazar, "Statistics and Data Visualisation with Python," Chapman and Hall/CRC, Jan. 31, 2023, doi: 10.1201/9781003160359.
- [19] M. Marudi, I. Ben-Gal, and G. Singer, "A decision tree-based method for ordinal classification problems," *IJSE Transactions*, vol. 56, no. 9, pp. 960–974, Jul. 2022, doi: 10.1080/24725854.2022.2081745.
- [20] J.-L. Goh, S.-B. Ho, and C.-H. Tan, "Weather-Based Arthritis Tracking: a mobile mechanism for Preventive Strategies," *Journal of Informatics and Web Engineering*, vol. 3, no. 1, pp. 210–225, Feb. 2024, doi: 10.33093/jiwe.2024.3.1.14.
- [21] N. Mrewa, A. M. Ramly, A. Amphawan, and T. K. Neo, "Optimizing Medical IoT Disaster Management with Data Compression," *Journal of Informatics and Web Engineering*, vol. 3, no. 1, pp. 55–66, Feb. 2024, doi: 10.33093/jiwe.2024.3.1.4.
- [22] J. Jayaram, Y. Kulkarni, L. V. Ganesh, Palanichamy Naveen, and Elham Abdulwahab Anaam, "Treatment Recommendation using BERT Personalization," *Journal of Informatics and Web Engineering*, vol. 3, no. 3, pp. 41–62, Oct. 2024, doi: 10.33093/jiwe.2024.3.3.3.
- [23] W.-X., Ong, S.-B., Ho, & C.-H., Tan, "Enhancing Migraine Management System through Weather Forecasting for a Better Daily Life," *Journal of Informatics and Web Engineering*, vol. 2, no. 2, pp. 201–217, Sept. 2023, DOI: 10.33093/jiwe.2023.2.2.15.
- [24] S.-K. Tan, S.-C. Chong, K.-K. Wee, and L.-Y. Chong, "Personalized Healthcare: A Comprehensive Approach for Symptom Diagnosis and Hospital Recommendations Using AI and Location Services," *Journal of Informatics and Web Engineering*, vol. 3, no. 1, pp. 117–135, Feb. 2024, doi: 10.33093/jiwe.2024.3.1.8.
- [25] S.-B., Ho, E.-Y., Chew, & C.-H., Tan, "Streamlining Dental Clinic Management for Effective Digitisation Productivity and Usability," *Journal of Informatics and Web Engineering*, vol. 3, no. 2, pp. 70–85, 2024, DOI: 10.33093/jiwe.2023.3.2.5.
- [26] R. Haque, S.-B. Ho, I. Chai, C.-W. Teoh, A. Abdullah, C.-H. Tan, & K. S. Dollmat, "Intelligent health informatics with personalisation in weather-based healthcare using machine learning," in *International Conference of Reliable Information and Communication Technology*, Cham: Springer International Publishing, pp. 29–40, Dec. 2020, doi: 10.1007/978-3-030-70713-2_4.
- [27] S.-B. Ho, S.-L. Chean, I. Chai, & C.-H. Tan, "Engineering meaningful computing education: programming learning experience model," in *2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, IEEE, pp. 925–929, 2019, doi: 10.1109/IEEM44572.2019.8978920.

- [28] S.-B. Ho, I. Chai, & C. H. Tan, "Leveraging framework documentation solutions for intermediate users in knowledge acquisition," *International Journal of Information Science*, vol. 3, no. 1, pp. 13-23, 2013.
- [29] S.-B. Ho, I. Chai, & C. H. Tan, "An empirical investigation of methods for teaching design patterns within object-oriented frameworks," *International Journal of Information Technology & Decision Making*, vol. 6, no. 4, pp. 701-722, 2007. doi: 10.1142/S021962200700271X.
- [30] I. Ibriwesh, S.-B. Ho, I. Chai, & C. H. Tan, "A controlled experiment on comparison of data perspectives for software requirements documentation," *Arabian Journal for Science and Engineering*, vol. 42, pp. 3175-3189, 2017. doi: 10.1007/s13369-017-2425-2.

BIOGRAPHIES OF AUTHORS

	<p>Chinwe Miracle Chituru is a PhD candidate in Computing at the Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia. Her research focuses on healthcare informatics, artificial intelligence, machine learning and quality information systems. Her specialization includes feature engineering, explainable AI and business intelligence. She can be contacted at email: reun.love36@gmail.com.</p>
	<p>Sin-Ban Ho is an Associate Professor in Multimedia University. He received his PhD degree in information technology from the Multimedia University, Cyberjaya, Malaysia, and teaches Computer Science at Multimedia University, Cyberjaya, Malaysia. His research interests include health informatics, empirical research approach, computer education, and patterns. He is a senior member of the IEEE, IEEE Computer Society, and IEEE Education Society. He can be contacted at email: sbho@mmu.edu.my.</p>
	<p>Ian Chai is a Principal Lecturer in Multimedia University. He received the B.S. degree in Computer Science in 1988, and the M.S. degree in Computer Science in 1991, both from the University of Kansas, Lawrence, USA. He worked in Connecticut and Germany, then received a Ph.D. in Computer Science from the University of Illinois at Urbana-Champaign, USA, in 2000. He now teaches Computer Science at Multimedia University in Cyberjaya, Malaysia. His research interests include computer science education, object-oriented design patterns, health informatics, and open-sourced software. He can be contacted at email: ianchai@mmu.edu.my.</p>